

# The Theoretical Astrophysical Observatory<sup>\*</sup>: Cloud-Based Mock Galaxy Catalogues

Maksym Bernyk<sup>1</sup>, Darren J. Croton<sup>1</sup>, Chiara Tonini<sup>1</sup>, Luke Hodkinson<sup>1</sup>, Amr H. Hassan<sup>1</sup>, Thibault Garel<sup>1,†</sup>, Alan R. Duffy<sup>2,3</sup>, Simon J. Mutch<sup>2,1</sup>, Gregory B. Poole<sup>2,1</sup>.

<sup>1</sup>*Centre for Astrophysics and Supercomputing, Swinburne University of Technology, Hawthorn, Victoria 3122, Australia*

<sup>2</sup>*School of Physics, University of Melbourne, Parkville, Victoria 3010, Australia*

<sup>3</sup>*ARC Centre of Excellence for All-Sky Astrophysics (CAASTRO)*

24 March 2014

## ABSTRACT

We introduce the Theoretical Astrophysical Observatory (TAO), an online virtual laboratory that houses mock observations of galaxy survey data. Such mocks have become an integral part of the modern analysis pipeline. However, building them requires an expert knowledge of galaxy modelling and simulation techniques, significant investment in software development, and access to high performance computing. These requirements make it difficult for a small research team or individual to quickly build a mock catalogue suited to their needs. To address this TAO offers access to multiple cosmological simulations and semi-analytic galaxy formation models from an intuitive and clean web interface. Results can be funnelled through science modules and sent to a dedicated supercomputer for further processing and manipulation. These modules include the ability to (1) construct custom observer light-cones from the simulation data cubes; (2) generate the stellar emission from star formation histories, apply dust extinction, and compute absolute and/or apparent magnitudes; and (3) produce mock images of the sky. All of TAO’s features can be accessed without any programming requirements. The modular nature of TAO opens it up for further expansion in the future.

**Key words:** galaxy formation, mock catalogue, light cone, online tools, cloud computing

## 1 INTRODUCTION

Astronomy has entered an era of survey science, thanks almost solely to instruments and telescopes that can sample unprecedented volumes of the cosmos with unparalleled sensitivity out to great distances. Riding this wave, the field of galaxy formation and evolution has grown to become one of the most active research areas in astrophysics. Progress feeds off progress, where increasingly detailed observations of galaxies are used to build new theories, that in turn lead to predictions, which can direct and be tested against new observations. As the instruments have gotten more sophisticated so has the amount of data they collect. Similarly, simulations of galaxies and the Universe have grown to keep pace.

It is thus not surprising that data access has become

a signature of this new era. Internet and cloud technologies allow scientists to store and retrieve large scientific datasets remotely. This is sometimes necessary since data volume and complexity often require resources beyond what is locally available. But even when not necessary it is frequently desired, as relocating storage and processing off-site reduces overheads, simplifies data management, and facilitates data sharing.

Two notable examples are the Sloan Digital Sky Survey (SDSS, Abazajian et al. 2003) “SkyServer”, which hosts imaging, spectra, spectroscopic and photometric data; and the German Astrophysical Virtual Observatory (GAVO, Lemson & Virgo Consortium 2006), which houses the Millennium Simulation theoretical data products. Both on-line repositories are accessible by means of the Structured Query Language (SQL), and due to their accessibility, both have vastly increased the scientific value of the data they hold through data re-use. There are many ways this benefit can be measured, arguably the most important being an increased

<sup>\*</sup> <https://tao.asvo.org.au/>

<sup>†</sup> Australian Research Council Super Science Fellow

number of scientific publications. Such publications come predominantly from researchers who had nothing to do with the original data production.

In large part due to this ease of access, the division between observer and theorist has faded somewhat. Observers now routinely use cutting edge theoretical models in their analysis, and theorists compare model predictions against observational data. This has all meant that the modern astronomer now routinely works across many traditional boundaries, often combining multiple disparate data products to undertake their science.

The focus of the present work is on access to theoretical survey data, such as cosmological-scale dark matter simulations and galaxy formation models. Many groups around the world are currently producing state-of-the-art theory products whose value to the community is immense. However access is often prohibitive, even when the authors are happy for others to use their work. Furthermore, comparing different simulations and models on an equal footing can be extremely problematic due to data size and transport barriers, data format differences, and complexity. This makes understanding how to correctly use the data challenging for the non-expert.

Access is but one part of the puzzle however. To be compared fairly to observations, simulations must typically be modified to look more like the data being compared against. This can include: mapping the simulation cube into an observed light-cone where distance also equates to time evolution in the simulation, calculating absolute and apparent magnitudes for model galaxies in select filters from their star formation and metallicity histories, and “observing” mock galaxies to generate images similar to that which would be collected by a CCD. All are non-trivial tasks that require great care to implement correctly.

Some effort has already gone in to producing such tools for the community. For example, the Mock Map Facility (MoMAF, Blaizot et al. 2005) allows a user to build mock galaxy catalogues using the GALICS semi-analytic model (Blaizot et al. 2004). In a similar fashion, the Millennium Run Observatory (Overzier et al. 2012) provides a powerful set of tools to build and analyse mock catalogues based on the Millennium Run suite of dark matter simulations (Springel 2005).

In this paper we present a new online tool – the Theoretical Astrophysical Observatory (TAO) – that aims to further address the problem of community data access and, more specifically, simplifies the process of building mock galaxy catalogues to more individualised specifications. TAO is a cloud-based<sup>1</sup> infrastructure that combines dark matter simulation data with semi-analytic galaxy formation and stellar population synthesis models. The usefulness of TAO is enhanced by a series of higher-level modules with which the user can modify the simulated data to better fit their science needs.

This paper is structured as follows: An introduction to TAO is presented in Section 2. In the subsequent sections we then describe the first four TAO science modules: the

basic galaxy and simulation selection tools (Section 3), the light-cone module (Section 4), the spectral energy distribution (SED) module (Section 5), and the mock image generation module (Section 6). We then explore usage cases in Section 7 that demonstrate the utility and functionality of TAO. Section 8 concludes with a summary.

For all results presented the cosmology of the simulation from which the result was drawn is assumed unless otherwise indicated, and we refer the reader to the associated reference for further details.

## 2 THE THEORETICAL ASTROPHYSICAL OBSERVATORY - AN OVERVIEW

The Theoretical Astrophysical Observatory (TAO) provides web access to mock extragalactic survey data generated using sophisticated semi-analytic galaxy formation models that are coupled to large N-body cosmological simulations. TAO is designed to be flexible, so that different simulations and models can be stored and accessed from a single location with a consistent data format. The interface for TAO is clean and built with simplicity in mind. All of TAO’s features require no programming knowledge to use, maximising accessibility to astronomers, be they observers or theorists.

A major feature of TAO is its ability to post-process the hosted data for different scientific applications. This is achieved through a number of science modules that can be chained in user-specified configurations, depending on the desired requirements of the astronomer and the module functionality. In addition, this modular design makes TAO readily expandable with new functionality in the future.

This paper presents the core science modules in TAO, which we summarise below and describe in more detail in the subsequent sections. These modules open up many science applications of value to astronomers.

- *Simulation data module.* This core module provides direct access to the original simulation and semi-analytic galaxy formation model data stored in the TAO SQL database. The user can specify the desired galaxy and dark matter halo properties to be retrieved at an epoch of interest from the simulation box (see Section 3).
- *Light-cone module.* This module remaps the spatial and temporal distribution of galaxies in the original simulation box on to that of the observer light-cone. The parameters of the cone are user configurable (see Section 4).
- *Spectral energy distribution module (SED).* This module retrieves the star formation and metallicity histories for each galaxy (either in the box or cone) from the TAO database and applies a user-selected stellar population synthesis model and dust model to produce individual galaxy spectra. These spectra are convolved with a set of filters in order to compute both apparent and absolute magnitudes (see Section 5).
- *Image module.* This module takes the output of both the light-cone and SED modules to construct user defined mock images. Images can be customised using a range of properties, such as sky area, depth, and a selected filter (see Section 6).

In Figure 1 we provide a broad overview of the TAO infrastructure. At the top level we define the connection between the account based user interface and user database.

<sup>1</sup> In this paper we define the “cloud” as Infrastructure as a Service (IaaS), which includes data storage and access, software infrastructure, and the use of supercomputer facilities.

In the middle level the various science modules are shown, as introduced above and which will be described in more detail in the subsequent sections. Both public and private data storage, containing the dark matter simulations, galaxy data, photometry etc., are shown in the lower level. On the back-end TAO is supported by a scalable database cluster hosted on the gSTAR supercomputer<sup>2</sup> at the Swinburne University of Technology. Each component of TAO affects the user experience and workflow, speed of mock data generation and retrieval, and the quality and utility of the final mock catalogue.

Users interact with TAO through a simple web form, where they can select a dark matter simulation, galaxy formation model, a box or cone geometry, and the associated parameters which define each. The desired galaxy and simulation properties to be included in the mock, including both absolute and apparent magnitudes in various filters, are all specified by the user. These selections are then passed to the back-end science modules via an XML parameter file, which can also be retrieved from the web interface for reference or later resubmission. Any required computations, e.g. to build a light-cone or set of SEDs, are then triggered on the gSTAR supercomputer for processing. The user is notified via email when their mock catalogue is completed, which may take from minutes to many hours depending on the size of the task. TAO also offers a choice of output formats, including CSV, HDF5 and FITS. The final mock catalogue can then be downloaded directly from the TAO website “History” tab to the user’s local machine. More advanced SQL/ADQL querying of the data is accessible through a VO Table Access Protocol (TAP) client, such as TOPCat.

In Figure 2 we show the TAO web interface, highlighting its minimalist yet functional design. The user interface design objective of TAO is to provide a simple portal that makes using the science modules easy and intuitive. In the spirit of reaching as many astronomers as possible, no specialist knowledge of SQL is required to use any part of TAO, keeping the barrier for access low.

TAO is part of the larger All-Sky Virtual Observatory (ASVO) project<sup>3</sup>, whose goal is to federate astronomy data and serve this to the wider community via the cloud. The ASVO constitutes a major infrastructure investment that links observational data with theoretical capabilities. It establishes a platform from which astronomers can optimally access and exploit the exponential growth in astronomical data volume in the coming decade. As a first release, the ASVO will include cloud access to the SkyMapper data archives (Keller et al. 2007), in addition to the simulated data provided through TAO. Ultimately it is expected that the ASVO will incorporate data at multiple wavelengths, including radio observations from the upcoming Australian Square Kilometre Array Pathfinder (ASKAP) telescope (Johnston et al. 2008).

### 3 THE GALAXY AND SIMULATION MODULE

This section will discuss the basics of the dark matter simulations and semi-analytic models that constitute the core

TAO data product. For completeness we feel it is important to provide such an overview give the central role such data plays in the functionality of TAO. More detail on the semi-analytic method can be found in Baugh (2006); Croton et al. (2006); Benson (2012). We also clarify some of the technical requirements of TAO to host this data. These include the requisite data format and minimum galaxy and halo properties required by the core and higher-level science modules.

#### 3.1 Dark matter simulations and large-scale structure

An efficient way to simulate the universe inside a supercomputer is to focus on the dominant mass distribution and its evolution. This usually involves running a collisionless N-body simulation in a volume that is large enough to be representative of the Universe as a whole, and provides a significant reduction in computational effort at fixed resolution compared with hydrodynamic galaxy formation simulations. Hydrodynamic effects are complicated and slow to compute numerically relative to the rather simple calculations required in a gravity-only simulation. Hence, gas and galaxies are often added later in post-processing using semi-analytic or other statistical techniques (see below).

As the universe evolves gravity pulls small structures together to assemble larger structures (i.e. hierarchical growth). Within the numerical simulation, such “halos” are typically identified using a Friends-of-Friends (FoF) algorithm (Davis et al. 1985; Springel et al. 2001; More et al. 2011), which detects gravitationally bound systems of particles and determines their properties. Structures within structures (i.e. sub-structures) can also be found using a variety of methods (e.g. Springel et al. 2005; Behroozi, Wechsler & Wu 2011). Such sub-structures are typically expected to host the smaller satellite galaxies and are subservient to the larger halo and central galaxy at the halo centre.

This information, calculated across all time-steps in a simulation for a particular object, defines its merger tree. The collection of such trees is then used as input to construct a galaxy formation model. An example halo merger tree from the Millennium Simulation Springel et al. (2005) is shown in Figure 3. Here, the top panel shows the tree itself for a  $1.9 \times 10^{13} M_{\odot}$  halo at  $z = 0$  (assuming  $h = 0.73$ ), while the corresponding mass growth history with time is shown in the lower panel.

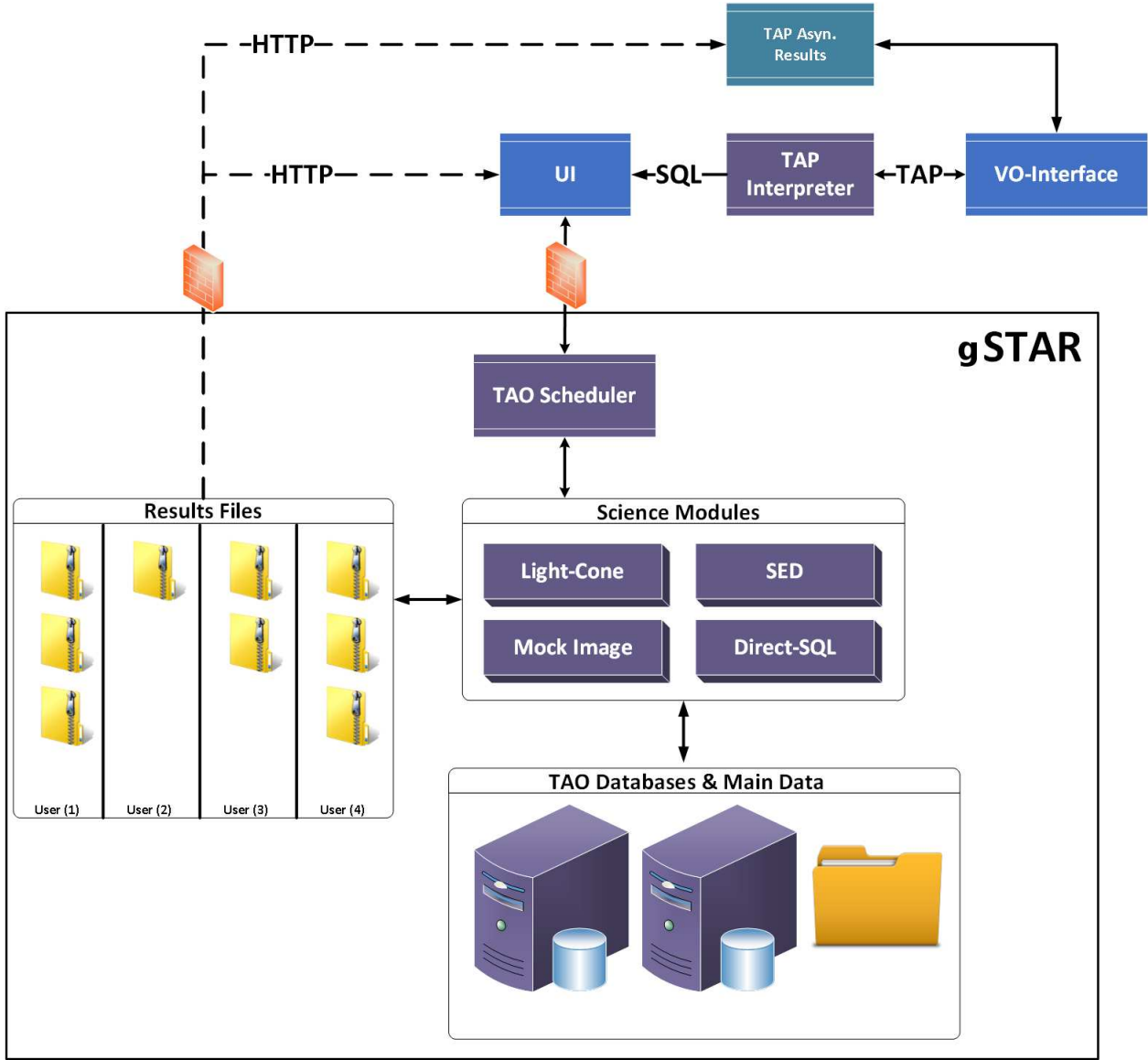
Simulations are run for different science goals, and each allow the exploration of different physics depending on how they were set-up. Therefore TAO provides the user with a choice of DM simulations, run with various sizes, mass resolutions and cosmological parameters.

#### 3.2 Modelling the evolution of galaxies

There are a number of ways to model the evolution of a galaxy inside a dark matter halo. For the higher-level science modules, TAO requires the galaxies in its database to have a minimum set of properties. As discussed below in Section 3.3, each galaxy record in the TAO database should contain information about its position, merger history, and star formation history. So, at a minimum, the methodology

<sup>2</sup> <http://www.astronomy.swin.edu.au/supercomputing/>

<sup>3</sup> <http://www.asvo.org.au/>



**Figure 1.** TAO architecture diagram. The main TAO database, science modules, and results storage space are located on the gSTAR supercomputer. The user interface and Table Access Protocol (TAP) server are hosted on a separate web server and accept job requests. Jobs are queued on the supercomputer via the jobs scheduler. Upon the job completion results are available for download over the internet.

used to generate the galaxy population must produce this information.

The best suited methodology for this purpose is that of semi-analytics (White & Frenk 1991). Semi-analytic models not only calculate all the required properties for galaxies, but also link their evolution over time using the halo merger trees to follow the growth histories. We provide a brief overview of the method here and refer the reader to Croton et al. (2006), Baugh (2006) and Benson (2012) for more information and details. Note that any other method that produces the minimum set of properties is also acceptable; we focus on semi-analytics here as it is the most common and nicely illustrates the requirements of the TAO system.

Semi-analytic galaxy evolution models take a dark matter halo as a kernel and then evolve its baryonic content us-

ing prescriptions that describe the phenomenology of each key process that affects the galaxy along its life cycle.

- (i) As a dark matter halo grows its potential well gets deeper and hence attracts baryons in the form of diffuse gas from the surrounding medium.
- (ii) This gas cools, conserving angular momentum as it falls to the centre of the halo and forms a rotationally supported disk.
- (iii) Within this flattened disk stars begin to form. A galaxy is born.
- (iv) Each episode of star formation results in a distribution of stellar masses sampled from the initial mass function.
- (v) The most massive stars are short-lived and explode as supernovae, injecting metals and energy back into the interstellar and intergalactic medium.

ASVO TAO (Beta)
New Catalogue
History
Admin
Documentation
Support

# New Catalogue

(Required Fields are marked with an asterisk)

Job Type
General Properties
Spectral Energy Distribution
Mock Image
Selection
Output format
Summary and submit

## Data Selection

Catalogue geometry \*

Light-Cone

Dark matter simulation \*

Millennium

Galaxy model \*

SAGE

Right Ascension Opening Angle (degrees) \*

10

Declination Opening Angle (degrees) \*

10

Redshift Min \*

0

Redshift Max \*

0.3

Estimated job size: 2%

☒ Unique  
☐ Random

Select the number of light-cones: \*

3

maximum is 3

## Output properties

Output properties \*

Available

Filter

**Galaxy Masses**  
 Bulge Stellar Mass  
 Cold Gas Mass  
 Hot Gas Mass  
 Ejected Gas Mass  
 Intracluster Stars Mass  
 Metals Total Stellar Mass  
 Metals Bulge Mass  
 Metals Cold Gas Mass  
 Metals Hot Gas Mass  
 Metals Ejected Gas Mass

>>  
 >  
 <  
 <<

**Selected**  
**Galaxy Masses**  
 Total Stellar Mass  
 Black Hole Mass  
**Positions & Velocities**  
 Right Ascension  
 Declination  
 Redshift (Cosmological)  
 Redshift (Observed)

Selected simulation details

**Millennium**

Cosmology: WMAP-1

Cosmological parameters:  $\Omega_m = 0.25$ ,  $\Omega_\Lambda = 0.75$ ,  $\Omega_b = 0.045$ ,  $\sigma_8 = 0.9$ ,  $h = 0.73$ ,  $n = 1$

Box size: 500 Mpc/h

Mass resolution:  $8.6 \times 10^8 M_{\text{sun}}/h$

Force resolution: 5 kpc/h

Paper: Springel et al. 2005

Selected galaxy model details

**SAGE**

Kind: semi-analytic model

Paper: Croton et al. 2006

Selected output property details

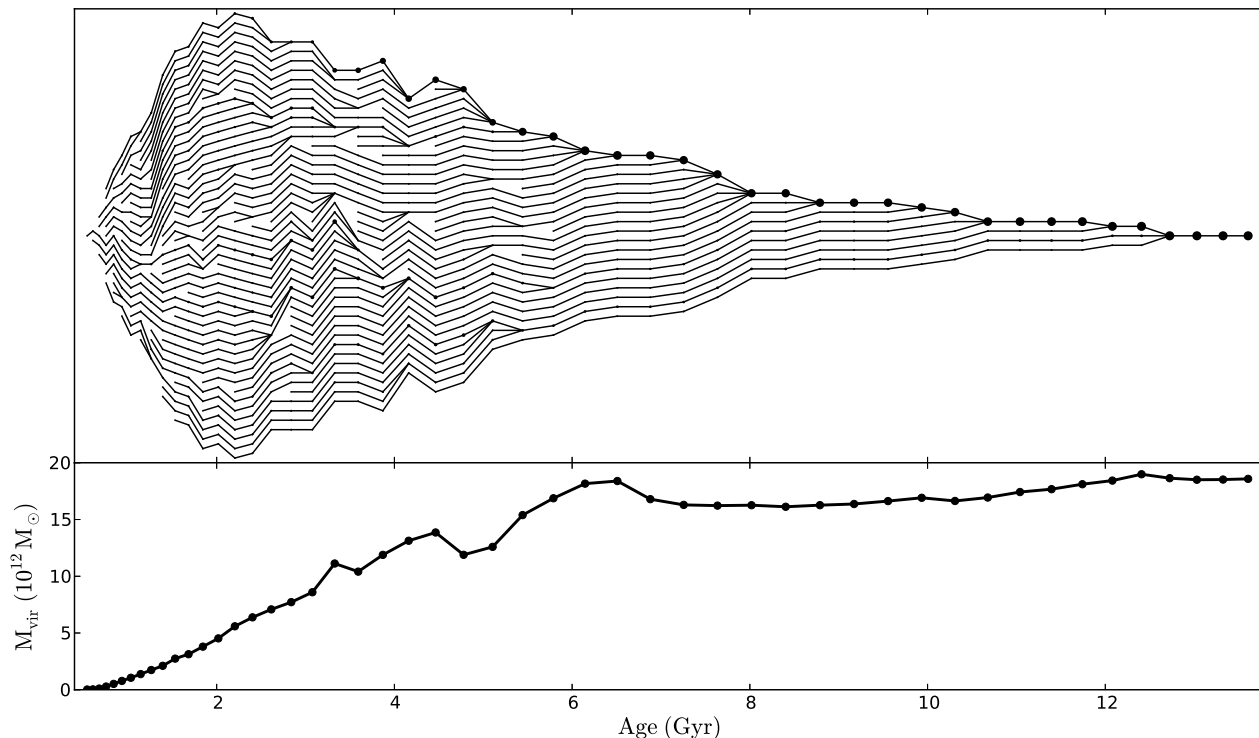
**Black Hole Mass** (10+10solMass/h)

Supermassive black hole mass

< Previous
Next >

**Figure 2.** The TAO web interface, showing how complex queries can be generated through a simple selection of galaxy and simulation properties, plus additional options to access the various science modules.

- (vi) Galaxies merge as hierarchical growth proceeds, resulting in morphological evolution and the birth of super-massive black holes.
- (vii) Supernova and super-massive black hole feedback can heat and/or remove gas from the disk and halo, gas that would otherwise contribute to the formation of the next generation of stars.
- (viii) Thus, an equilibrium state is established as the galaxy goes through a cycle of stellar birth, feedback, gas heating/removal and star formation suppression.
- The above processes provide us with a history of galaxy properties required to produce a mock catalogue. The simulations and models used in TAO are often large, tracking



**Figure 3.** A dark matter halo merger tree drawn from the Millennium Simulation (top) and its mass evolution (bottom) with time. This halo has a final mass of  $1.9 \times 10^{13} M_{\odot}$  (assuming  $h = 0.73$ ) and would be typical of a group sized system in the real Universe.

many tens of millions of halos (and hence galaxies). It is from the properties of such large catalogues of simulated galaxies that the output of TAO is derived.

### 3.3 TAO input data

Each simulation and galaxy model will potentially contain a vast array of properties of interest to astronomers. TAO groups its data for each halo–galaxy pair into the following categories:

- Baryonic masses, such as stellar mass, cold gas mass, and hot halo gas. Metals for each baryonic component are included in this category.
- Galaxy properties, such as the star formation rate, disk scale radius, and cooling and AGN heating rates.
- Halo properties, such as halo mass and other virial properties, spin and velocity dispersion.
- Positions & velocities common for both halos and galaxies. These are always given in the co-moving Cartesian coordinates of the simulation box.
- Simulation properties. This includes the temporal snapshot number and any galaxy and halo IDs.

In Table 1 we provide a summary of the minimum property requirements that TAO needs to operate, broken down by science module.

To execute the science modules (and in particular, the SED module described in Section 5) TAO must be able to walk each halo merger tree from any point in its history. Hence, for each object we require pointers that identify both the previous time-step progenitor and that link the sequence of subhalos in a FoF halo at a given time-step in order of

decreasing mass. We adopt the SUBFIND system of pointers, assuming they are stored in depth-first order for each merger tree, as illustrated by Figure 11 of the Supplemental Material in Springel et al. (2005).

To build a mock light-cone the data must also contain Cartesian coordinates for each halo–galaxy pair from the original simulation box; these are then converted to angular coordinates and redshifts using methods described in Section 4. Furthermore, conversion to redshift space requires Cartesian velocities (i.e. proper motion of the galaxies).

To calculate magnitudes the SED module must be able to extract star formation and metallicity histories for each galaxy. Also, the initial mass function (IMF) assumed when calibrating the model is required (e.g., this typically constrains the recycling fraction of mass returned to the interstellar medium from stellar winds).

#### 3.3.1 Units

We conclude this section by defining the unit and Hubble constant conventions assumed in TAO. The use of little  $h$  in particular can differ significantly between the theory and observational communities (and even within a community) and care must always be taken (Croton 2013).

- Masses adopt  $h^{-1} M_{\odot}$ .
- Positions adopt  $h^{-1} \text{Mpc}$  and are always in co-moving coordinates.
- Sizes adopt  $h^{-1} \text{Mpc}$  and are always in physical coordinates.
- Velocities adopt  $\text{km s}^{-1}$  and are in co-moving coordinates.
- Star formation rates adopt  $M_{\odot} \text{yr}^{-1}$ .
- Heating and cooling rates are in units of  $\log_{10} \text{erg s}^{-1}$ .

Module	Minimum information required
<b>Simulation &amp; Model (§3)</b>	<ul style="list-style-type: none"> <li>• A list of the simulation snapshot time-steps. The box size, particle mass resolution and assumed cosmological parameters.</li> </ul>
<b>Light-Cone (§4)</b>	<ul style="list-style-type: none"> <li>• Co-moving Cartesian coordinates (<math>x</math>, <math>y</math>, <math>z</math>) and velocities (<math>v_x</math>, <math>v_y</math>, <math>v_z</math>) for each halo/galaxy at each time-step.</li> <li>• IDs that associate subhalos/satellites with their parent central halo/galaxy.</li> </ul>
<b>SED (§5)</b>	<ul style="list-style-type: none"> <li>• Indices connecting the history of each halo/galaxy across time through the galaxy merger tree (see Section 3.3).</li> <li>• Galaxy star formation rate and metallicity at each time-step.</li> <li>• The initial mass function assumed by the galaxy model.</li> </ul>
<b>Mock Images (§6)</b>	<ul style="list-style-type: none"> <li>• A measure of galaxy morphology, such as independent galaxy disk and bulge stellar masses.</li> <li>• The output from the light-cone and the SED modules.</li> </ul>

**Table 1.** Data requirements for the TAO modules.

- Photometric magnitudes are calculated assuming the  $h$  value of the dark matter simulation used to produce the mock catalogue.

## 4 THE LIGHT-CONE MODULE

N-body simulations of the type discussed in the previous section typically adopt a periodic three-dimensional box geometry. This is in contrast to the geometry observed with a telescope, where galaxies are seen strung out along the observer’s light-cone. Converting between a box and a light-cone for the purpose of building a mock catalogue is a mechanical yet non-trivial task. In TAO, this is taken care of by the light-cone module. To explain cone construction, we start with some basic concepts and build up to the more sophisticated method used by TAO.

### 4.1 Basic light-cone construction

To build a basic light-cone we place an observer at one of the corners of the simulation box and have them “look out” at the model galaxy distribution. We do this by remapping the Cartesian coordinates of each galaxy into their angular positions in right ascension (RA), declination (Dec) and radial distance ( $d$ ). This operation defines the basic cone geometry in real-space.

$$\begin{cases} d = \sqrt{x^2 + y^2 + z^2} \\ \text{RA} = \arctan\left(\frac{x}{y}\right) \\ \text{Dec} = \arcsin\left(\frac{z}{d}\right) \end{cases} \quad (1)$$

Here  $x$ ,  $y$ , and  $z$  are the coordinates along the principal axes of the simulation box.

However, TAO has the capacity to generate cones in either real-space or redshift-space (i.e. with or without line-of-sight peculiar velocities factored into the radial distance). For redshift-space cones, each galaxy distance is instead calculated using

$$d_{\text{rs}} = d + \frac{1}{H_0} \left( \frac{x}{d} v_x + \frac{y}{d} v_y + \frac{z}{d} v_z \right), \quad (2)$$

where  $v_x$ ,  $v_y$ , and  $v_z$  are the velocity coordinates for the galaxy in km/s, and  $H_0$  is the Hubble constant. The above has assumed co-moving coordinates for all positions and distances, which is standard in modern simulations.

For TAO to provide a redshift for each galaxy on the cone it is necessary to invert the radial (co-moving) distance using the distance–redshift relation for the given simulation cosmology (Hogg 1999). This is given by

$$D_C = D_H \int_0^z \frac{dz'}{E(z')}, \quad (3)$$

where  $D_H = c/H_0$  is the Hubble distance,  $c$  is the speed of light,  $z$  is now redshift, and the expansion factor  $E(z)$  is defined as

$$E(z) \equiv \sqrt{\Omega_M(1+z)^3 + \Omega_k(1+z)^2 + \Omega_\Lambda}. \quad (4)$$

$\Omega_M$ ,  $\Omega_k$ , and  $\Omega_\Lambda$  are the matter density, curvature, and cosmological constants, respectively. To obtain the real-space or redshift-space redshift, TAO simply uses either  $d$  or  $d_{\text{rs}}$  for  $D_C$  when inverting Equation 3, respectively.

### 4.2 Expanding the cone beyond the box

A problem with the above cone construction becomes apparent when building cones that are deeper in radial extent than the box from which the cone is cut. However, there are a number of ways to deal with this. All of them rely on the fact that most modern simulations are run assuming periodic boundary conditions, meaning that each side of the box connects seamlessly with its opposite side. Below we will assume the box is replicated in all directions, as required to construct such an extended cone. There are now two cases to consider.

#### 4.2.1 Unique cones

To construct an extended cone we replicate the box in the desired direction and continue the cone construction into this ‘new’ box. To ensure the cone is unique – i.e. any given galaxy in the box along any point in its history is only featured in the cone at most once – we carefully select the viewing angle so that the cone never overlaps with itself as it extends (Carlson & White 2010). This method limits the size of the cone to have a total volume smaller than the box from which the cone is cut; typically much smaller.

Within TAO, to optimally select the unique cone viewing angle we need to satisfy the following condition: the light-cone intersection with the furthest and the closest surfaces of the last replicated simulation box must not overlap when both are projected on to a plane of either surface. If this condition is satisfied for the last replicated simulation box, then it will also be satisfied for the rest of the boxes

because the light-cone gets narrower closer to the observer, thus making larger gaps between the light-cone path inside the one original simulation box. This is illustrated in Figure 4.

The condition for a unique cone can be written as

$$\left(d - \frac{b}{\cos(\alpha + \beta)}\right) \sin(\alpha + \beta) \geq d \sin \alpha, \quad (5)$$

where  $d$  is the depth of the light-cone,  $b$  is the side length of the simulation box,  $\alpha$  is the viewing angle of the cone from the origin in the declination plane,  $\beta$  is the cone opening angle in the declination plane, and  $\gamma$  is the cone opening angle in the right ascension plane. Furthermore,  $\gamma$  requires an additional condition be satisfied:

$$d \cos(\gamma) \leq b. \quad (6)$$

For the cone to make efficient utilisation of the volume, conditions 5 and 6 dictate that

$$\frac{V_{\text{cone}}}{V_{\text{min}}} = \frac{\sin \beta \cos \frac{\gamma}{2}}{3 (\sin(\alpha + \beta) - \sin \alpha) \cos(\alpha + \beta)}, \quad (7)$$

where  $V_{\text{cone}}$  is the volume of the cone, and  $V_{\text{min}}$  is the minimum rectangular volume required from the simulation box to fully contain the unique cone.

#### 4.2.2 Random cones

More typically, however, the volume of the desired cone is larger than the volume of the simulation cube. In this case one can build a random cone. Although such cones result in the replication of structure, any periodicity can be mitigated somewhat by using randomisation techniques (e.g. Blaizot et al. 2005) which produce a more realistic light-cone with pseudo-unique structure (i.e. repeated but non-periodic).

To remove the appearance of periodically repeating structures three randomisation transformations are applied within the TAO light-cone module: random rotation, mirroring, and translation of each repeated simulation cube.

**Rotation:** The rotation matrix of the principal axis  $x$ ,  $y$  and  $z$  is given by

$$\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = \begin{bmatrix} \hat{\mathbf{u}}_x & \hat{\mathbf{v}}_x & \hat{\mathbf{w}}_x \\ \hat{\mathbf{u}}_y & \hat{\mathbf{v}}_y & \hat{\mathbf{w}}_y \\ \hat{\mathbf{u}}_z & \hat{\mathbf{v}}_z & \hat{\mathbf{w}}_z \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} \quad (8)$$

$$\begin{cases} \hat{\mathbf{u}}_x = \cos \theta \cos \psi \\ \hat{\mathbf{u}}_y = \cos \theta \sin \psi \\ \hat{\mathbf{u}}_z = -\sin \theta \end{cases} \quad (9)$$

$$\begin{cases} \hat{\mathbf{v}}_x = -\cos \phi \sin \psi + \sin \phi \sin \theta \cos \psi \\ \hat{\mathbf{v}}_y = \cos \phi \cos \psi + \sin \phi \sin \theta \sin \psi \\ \hat{\mathbf{v}}_z = \sin \phi \cos \theta \end{cases} \quad (10)$$

$$\begin{cases} \hat{\mathbf{w}}_x = \sin \phi \sin \psi + \cos \phi \sin \theta \cos \psi \\ \hat{\mathbf{w}}_y = -\sin \phi \cos \psi + \cos \phi \sin \theta \sin \psi \\ \hat{\mathbf{w}}_z = \cos \phi \cos \theta \end{cases} \quad (11)$$

where  $\phi$ ,  $\psi$  and  $\theta$  are Euler angles. For the sake of performance TAO randomly takes values 0, 90, 180 or 270 degrees.

**Mirroring:** Mirroring of the simulation volume can be simply achieved by changing an axis direction. It should be noted that inversion of all of the principal axis in combination with rotation may result in the original positions in the simulation cube, so these combinations must be excluded from the randomisation routine.

**Translation:** To translate the cube we cut the simulation box at a random position along one axis and move the sliced volume before this position to the end of the simulation box along of the same axis. This operation doesn't affect continuous structure in the box because of the periodic boundary condition in the original simulation.

To illustrate the effects of replication and randomisation we construct two mock light-cones built with TAO using the milli-Millennium Simulation (Springel et al. 2005) having box side-length  $62.5 h^{-1} \text{Mpc}$ . This is shown in Figure 5. The first assumes straight replication of the simulation box (left panel), while the second applies the above random rotations, shifting and mirroring (right panel).

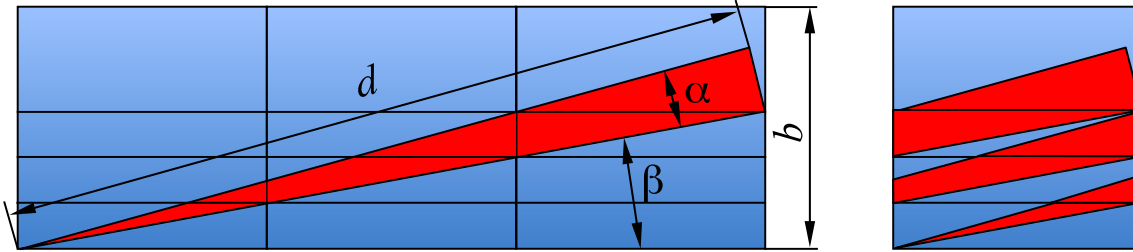
The differences between the left and right panels of Figure 5 are striking. At a bare minimum, given the visual nature of astronomical research clearly non-periodic mock catalogues are desirable. But more importantly, randomisation removes the possibility of unintended replicated features creeping into statistical applications of the mock. We emphasise however that spatially-dependent results should never be drawn from mocks on scale-lengths larger than the simulation box itself.

To quantify the effect of such operations on the spatial distribution of galaxies along the cone, in Figure 6 we plot the real-space 2-point correlation function for galaxies drawn from TAO that are more massive than  $10^{10} h^{-1} \text{M}_{\odot}$ . We first do this for original galaxy positions in the the Bolshoi simulation box (Klypin, Trujillo-Gomez & Primack 2011), which has a side-length of  $250 h^{-1} \text{Mpc}$  (solid blue line). We then compute the clustering in 20 randomised cones cut from the same (full) simulation box, each with an area of  $30 \times 30$  degrees on the sky covering redshift  $0 < z < 0.2$  (up to  $576 h^{-1} \text{Mpc}$  depth along the line-of-sight, dashed red line). This selection ensures that volumes are similar for both cones and the box. Each of the randomised cones span 6 replicated boxes and therefore includes 6 displaced by randomisation concatenations. The  $3\sigma$  scatter in the clustering results are used as a measure of the clustering uncertainty in the generated cones sample, shown by the error bars. Figure 6 shows consistent behaviour between the original galaxy distribution and randomised cones across the range of scales plotted.

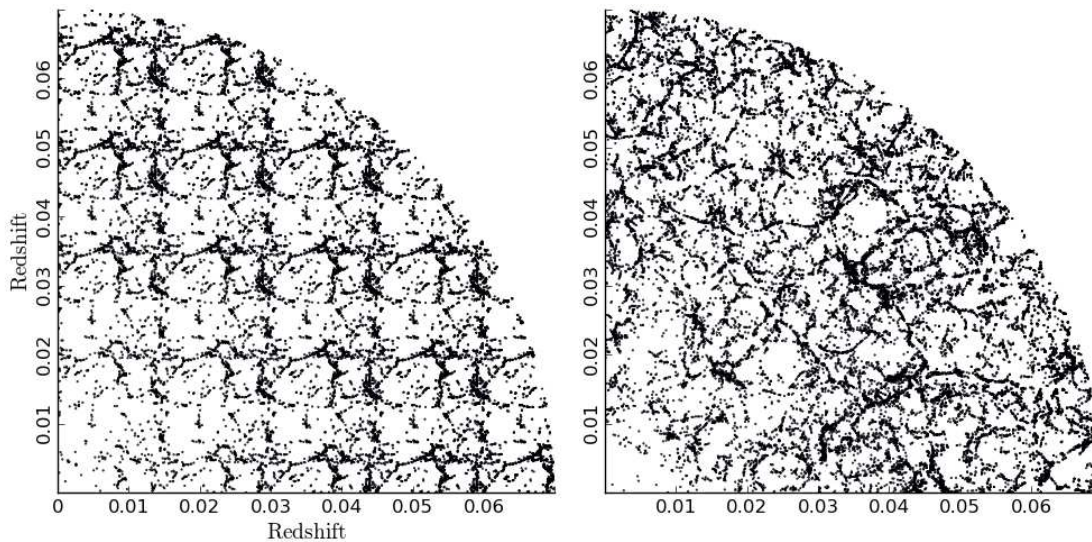
#### 4.3 Time evolution along the cone

As we progressively move back through the light-cone away from the point of observation we see galaxies from earlier and earlier epochs, due to the time required for the light from each galaxy to cover the distance. Therefore, to construct an accurate cone we need to not only worry about the spatial distribution of the galaxies in it, but their evolution as well. Remembering that a galaxy's radial distance can be directly mapped to a cosmic time, we walk the history of each cone galaxy in the TAO database until we reach the closest simulation output corresponding to the age of the Universe at that cone position. The galaxy at that point in





**Figure 4.** Selecting a unique path for a light-cone through a simulation cube. The left side shows the replicated boxes for an optimal unique light-cone. The right side is the same structure but collapsed into a single simulation cube.



**Figure 5.** Two volume limited light-cones built from a simulation with a box side-length of  $62.5 h^{-1} \text{ Mpc}$ . The left cone shows the result of standard box replication, while the right cone includes the random rotation, shifting and mirroring techniques employed by TAO to minimise any periodicity, as discussed in Section 4.2.2.

its evolution is then selected for the cone, and so on for all cone galaxies.

By doing this we also reduce the consequences of structure replication during the cone construction processes. Not only will repeated large-scale structures be seen from different orientations due to the random shuffling described above, but they are likely to be earlier (or later) versions of their duplicates, perhaps appearing differently depending on the growth history of each halo–galaxy system.

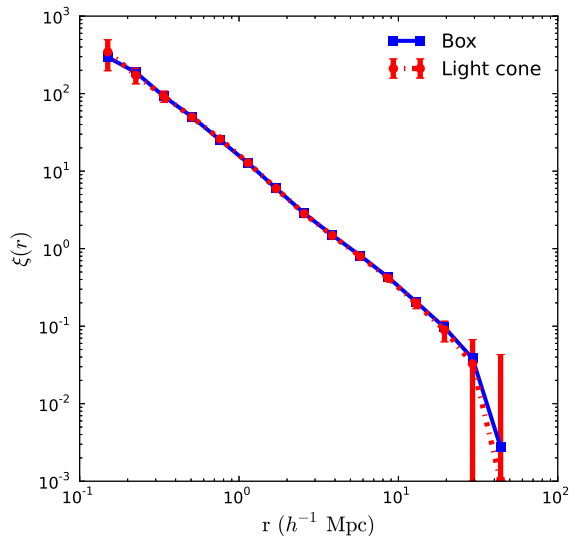
There is an additional complication here that is quite subtle yet important, as illustrated in Figure 7. A key part of the light-cone production is to make sure satellite galaxies are always connected to their parent dark matter halo. It often happens that a dark matter halo has its position in the cone very close to the border between different simulation time-steps. If so, satellite galaxies in that halo may be split in time across the boundary, and then also be displaced by the randomisation part of the algorithm. In order to provide structural consistency, when building the light-cone we group galaxies (centrals and satellites) by their parent halo association as given in the TAO database. Based on the position of the halo centre we then insert these as a unit into

the cone, even if the unit spatially crosses the time boundary at any point. This ensures that entire galaxy–halo structures at the same time-step are selected for the cone.

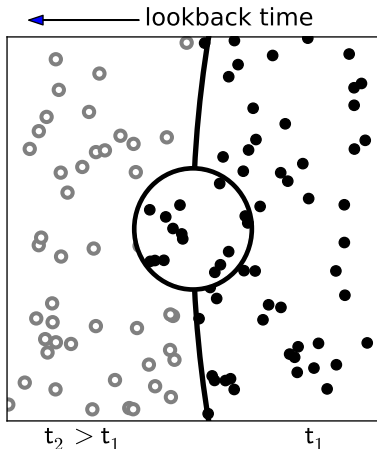
## 5 THE SPECTRAL ENERGY DISTRIBUTION MODULE

The next crucial step to building mock galaxies is to model their stellar emission, which further enables the translation of theoretical quantities into mock observables. The light emitted by a galaxy is the direct outcome of the formation and evolution of stars, which is regulated by all the physical mechanisms involved in galaxy formation.

From a practical point-of-view, modelling galaxy emission can be separated from the rest of the galaxy formation model. Hence, in TAO, determining the light emission is performed as a post-processing step in the Spectral Energy Distribution module applied to the TAO galaxy data. The post-processing link is made possible using the model’s prediction for the star formation rate history and metallicity history of each galaxy.



**Figure 6.** Galaxy correlation function measured in the original Bolshoi simulation box (blue solid line), and in a randomised mock light-cone constructed from the box using the procedures described in Section 4 (red dashed line).



**Figure 7.** Subhalos from the same parent dark matter halo on a border between two redshift zones are kept together. The solid dots show the galaxy positions from the previous time-step in the simulation, whereas the open circles show galaxies from the next time-step. The large circle in the middle defines the region occupied by one particular dark matter halo, which we preserve from splitting when applying time evolution across the cone galaxies.

### 5.1 Stellar population synthesis models

Galaxy light is the superposition of the emission of all the stars in the galaxy. A galaxy is composed of a series of single stellar populations (SSPs), i.e. ensembles of stars formed in single episodes with the same age and metallicity. The SSPs that compose a galaxy either originate in the galaxy itself through star formation, or are accreted from satellite galaxies. The emission of every SSP has to be modelled and added to the total galaxy light. Obviously, as the stars in the galaxy age their emission changes with time, and the model needs to take this time-dependance into account.

The tools used to accomplish this are stellar population synthesis (SPS) models, which are libraries of spectra of single stellar populations built on a grid of ages and metallicities, assuming a particular initial mass function (IMF) (e.g. Bruzual & Charlot 2003; Maraston 2005; Conroy, Gunn & White 2009). In order to model the galaxy emission as a function of time, at every time-step in the galaxy model we keep track of all the single stellar populations in the galaxy, then assign them the corresponding emission based on their age and metallicity. We then sum over all populations to obtain the total galaxy light, i.e. its spectral energy distribution (SED). At this stage the complicating contribution of dust extinction and emission must also be included.

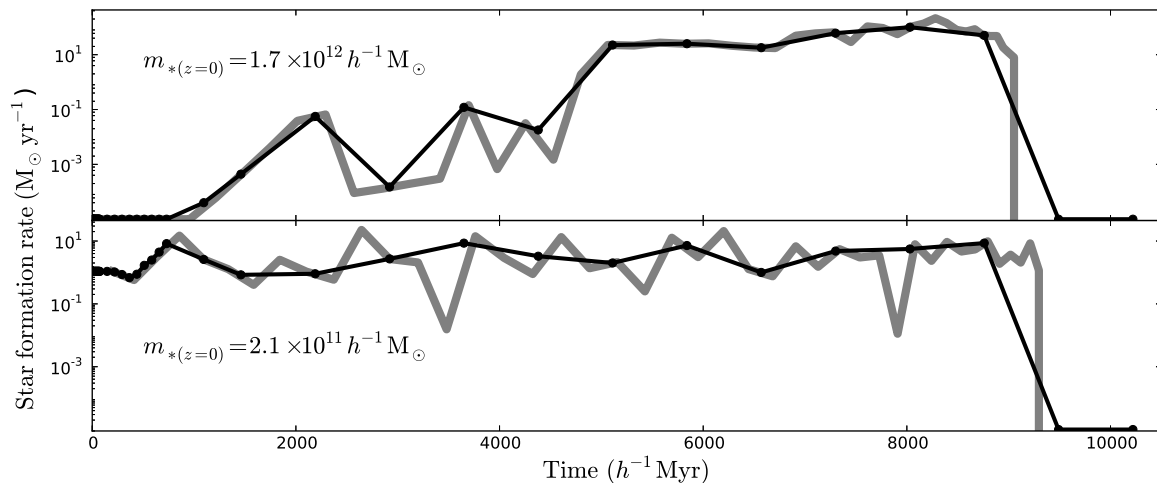
Semi-analytic models are now taking advantage of such methods to generate galaxy light (Hatton et al. 2003; Tonini et al. 2009, 2010; Henriques et al. 2011), which add a level of sophistication and flexibility. The use of spectra, as opposed to magnitude tables (De Lucia, Kauffmann & White 2004; Croton et al. 2006; Baugh 2006; Bower et al. 2006) brings a number of advantages: 1) an increased precision due to the linear additive nature of luminosity *vs* the logarithmic behaviour of magnitudes; 2) an increased accuracy for the determination of observed magnitudes, which are obtained by integration on redshifted SEDs rather than through theoretical *k*-corrections that rely on toy-model spectra; 3) an enormous flexibility, introduced when SEDs are modelled in post-processing. Producing galaxy spectra allows us to separate the photometric calculations from the semi-analytic model itself, removing the need to re-run the galaxy model every time we want to change the photometry specifications. This can include different initial mass functions (IMF), dust models, filter sets, mock observational errors, and telescope or survey-specific effects.

### 5.2 Galaxy star formation histories

To calculate a galaxy SED within TAO we require its star formation history, defined as the stellar populations present in the galaxy at the time of observation  $\tau_0$ , characterised by stellar mass, age, and metallicity. These populations include stars formed in the galaxy itself and those that have been accreted from satellites along the merger tree<sup>4</sup>.

We build a two-dimensional age and metallicity grid and collapse onto it the stars formed across the entire tree up to the point of observation (i.e. up to the point  $\tau_0$ , when we observe the galaxy on the cone). Here, each bin represents a single stellar population of a given age and metallicity, and from the SPS libraries we select the corresponding spectrum and weight it by the stellar mass formed. Each of these spectra are then added to the total galaxy light to produce the final SED.

<sup>4</sup> As an aside, the capacity to record the star formation history of a galaxy in its actual merger tree is an undeniable advantage of semi-analytic models over many other models techniques: without the merger tree information toy-models of galaxy evolution can not account for the complexity of the hierarchical nature of the galaxy assembly, potentially introducing significant biases (Tonini et al. 2012).



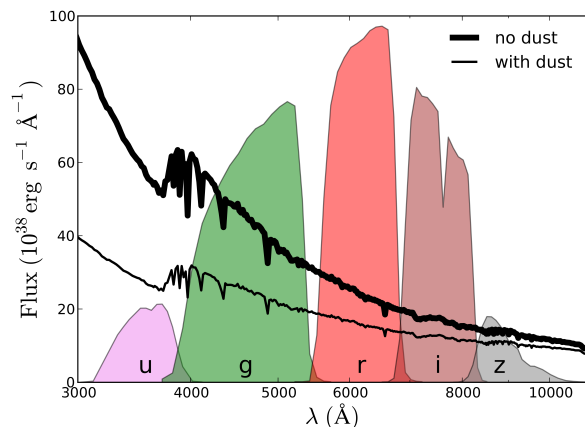
**Figure 8.** Star formation histories of two galaxies in the TAO database: a large elliptical galaxy (top) and a large Milky Way-type spiral galaxy (bottom). For each, the grey thick line shows the original star formation history from the semi-analytic model on the time grid of the dark matter simulation, whereas the thin black line with points indicates the interpolated star formation history on the grid required by the SED module in TAO.

The challenge in the production of the star formation history is the grid itself. SPS models are built to follow the vastly different speeds of stellar evolution. For example, the emission of a young stellar population is dominated by massive stars and changes on time-scales of  $\sim 1$  Myr, while an old stellar population emits steadily on time-scales of  $\sim 1$  Gyr, with emission dominated by less massive main-sequence stars. The star formation history grid needs to provide information on corresponding time-scales in order to produce realistic galaxy SEDs. In particular, the ultra-violet and optical part of the SED is heavily (if not entirely) determined by young stellar populations, an issue that becomes extremely important at high redshifts.

To this end, the star formation histories of every galaxy in the light-cone must be written by TAO onto a time-varying grid, anchored at the time of observation  $\tau_0$ , being finely spaced (steps of 1 Myr) for young ages near  $\tau_0$  and more sparsely spaced towards older ages. Unfortunately, the intrinsic semi-analytic model time grid (the grid on which the model calculates its physics) is not typically spaced like this. For example, the average time step of a model built using the Millennium Simulation is of the order of  $\sim 300$  Myr. Furthermore, refining a galaxy model grid to the level of precision required by SPS models is not usually practical due to the huge amount of data storage that would be required.

To make up for the loss of information on smaller time-scales we spread the time-weighted stellar mass produced in each larger semi-analytic model output step onto the fine SPS time grid close to  $\tau_0$ . Note that this is equivalent to assuming that in each simulation time bin we have a constant star formation rate. At the other end of the SPS time grid the opposite is done, where the mass produced across multiple model time-steps are collected and re-binned to fill the coarser SPS grid, accounting for the slow evolution of the old stellar populations.

In Figure 8 we plot two example star formation histories from a semi-analytic model in TAO, one for a spiral



**Figure 9.** Synthetic spectra of a galaxy from TAO at  $z = 0$  that has a star formation rate of  $15 M_{\odot} \text{ yr}^{-1}$  and stellar mass  $2.1 \times 10^{10} h^{-1} M_{\odot}$ . Here the Maraston (2005) SSP model has been assumed. The thick line shows the spectra without dust extinction, while the thin line is the same but with the dust model from Tonini et al. (2012) applied. Over-plotted are the SDSS  $u$ ,  $g$ ,  $r$ ,  $i$ ,  $z$  transmission functions using an arbitrary y-axis scale.

galaxy and one for an elliptical. In both panels, the *grey line* represents the galaxy star formation rate binned using the original simulation time grid, summed over all the branches in the merger tree, while the *black line* represents the star formation history interpolated over the SPS time grid used in TAO. Notice how, for recent times (i.e. close to  $\tau_0$ ), the spacing of the SPS grid is very fine, and the interpolation recovers the SAM star formation history exactly, while for older ages the spacing becomes sparser, providing an approximated reconstruction.

### 5.3 Galaxy spectro-photometric properties

TAO produces galaxy photometry after determining the SEDs. Magnitudes are calculated by convolving each galaxy

spectrum with a set of filter transmission functions, which include filters from many of the most commonly used instruments and surveys. TAO interpolates filters and spectra on a variable wavelength grid, so that the resolution of the integration is constant no matter the wavelength extension of the filter function.

The user is given the choice of both absolute and apparent magnitudes for each filter. In the case of absolute magnitudes, fluxes are calculated from the total luminosity on a sphere of radius  $R = 10\text{pc}$ . For apparent magnitudes, the flux is calculated on a sphere of radius equal to the luminosity distance corresponding to the redshift of the galaxy. The spectrum is dimmed and stretched in wavelength according to the redshift, and then convolved with each selected filter function. Notice that this operation is exact, as opposed to the approximation of using a K-correction, as both the rest-frame and observed (stretched) spectra are known.

In Figure 9 we show an example SED from a  $z = 0$  star forming galaxy in TAO (thin line), where the Maraston (2005) SPS was chosen. It is evident that recent and more intense star formation increases the stellar emission in the UV and optical wavelengths, and the young stellar populations dominate the bolometric luminosity. Also shown is the effect of including the contribution of dust extinction (in the UV/optical) and emission (mid-to-far infrared; thin line). The inclusion of dust in the theoretical spectra is available to the user as an option within the TAO SED science module which we now discuss.

## 5.4 Dust

Dust emission and extinction plays a fundamental role in shaping the galaxy SEDs, especially in cases of significant star formation, which is particularly relevant at high redshifts.

In TAO, we provide dust modelling as a separate step and allow the user to choose whether to apply dust to the galaxy SED or not. Two popular models are available in TAO and are described below.

### 5.4.1 Slab model

In the slab model, stars and dust are assumed to be homogeneously distributed in an infinite plane-parallel slab with the same vertical scale. The dust-attenuated luminosity at wavelength  $\lambda$ ,  $L_\lambda^{\text{obs}}$ , is given by:

$$L_\lambda^{\text{obs}} = L_\lambda^{\text{intr}} \frac{1 - e^{-\tau_\lambda^{\text{eff}} \sec i}}{\tau_\lambda^{\text{eff}} \sec i}, \quad (12)$$

where  $L_\lambda^{\text{intr}}$  is the intrinsic luminosity of the disc and  $i$  is its inclination angle. We define the effective dust opacity as  $\tau_\lambda^{\text{eff}} = (1 - \omega_\lambda)^{1/2} (1 + z)^{-1/2} \tau_\lambda$ , where  $\tau_\lambda$  is the face-on dust opacity. Following Devriendt, Guiderdoni & Sadat (1999)  $\tau_\lambda$  is expressed as a function of the neutral hydrogen column density of the disc,  $N_{\text{H}}$ :

$$\tau_\lambda = \left( \frac{A_\lambda}{A_V} \right)_{Z_\odot} \left( \frac{Z}{Z_\odot} \right)^s \left( \frac{N_{\text{H}}}{2.1 \times 10^{21} \text{ atoms cm}^{-2}} \right). \quad (13)$$

Here  $(A_\lambda/A_V)_{Z_\odot}$  is the extinction curve for solar metallicity  $Z_\odot$  (see Mathis, Mezger & Panagia 1983) and varies with gas metallicity  $Z$  and wavelength such that  $s = 1.35$  for  $\lambda > 2000\text{\AA}$ , and  $s = 1.6$  for  $\lambda < 2000\text{\AA}$  (see Guiderdoni & Rocca-Volmerange 1987 for more details).

The first term of Eq. 13,  $(1 - \omega_\lambda)^{1/2}$ , accounts for scattering effects, where  $\omega_\lambda$  is the albedo. The second term of Eq. 13, often used in semi-analytic models (e.g. Kitzbichler & White 2007, Garel et al. 2012), introduces an additional scaling of the dust-to-gas ratio with redshift which is in broad agreement with observational trends seen in high redshift galaxies, e.g. Reddy et al. (2006).

### 5.4.2 Calzetti prescription

The dust content of a galaxy can be parameterised with the colour excess  $E(B - V)$ , which is defined as a re-normalisation of the spectrum. Physically, dust content is associated with the presence of supernovae Type II, which are the main contributors to the metals that constitute the dust grains. Such grains are short-lived (with a life-span of the order of  $\sim 10 - 100$  Myr), so it is sensible to associate the dust content with the instantaneous star formation rate in the galaxy (see Tonini et al. 2011),

$$E(B - V) = R_{\text{dust}} A \cdot \left( \frac{\dot{m}_*}{\dot{m}_{*,0}} \right)^\gamma + B. \quad (14)$$

Here,  $A = (e^3 - e^{-2})^{-1}$ ,  $B = -Ae^{-2}$ ,  $\dot{m}_{*,0} = 1.479 M_\odot \text{y}^{-1}$ , and  $\gamma = 0.4343$  are values obtained with a calibration of the GOODS sample discussed in Daddi et al. (2007) and the Andromeda galaxy. We use a Calzetti extinction curve, which produces absorption blue-wards of the Johnston K band, and re-emission red-wards (see Calzetti 1997 and Calzetti 2001).

## 6 THE MOCK IMAGE MODULE

In addition to the spectral energy distribution of individual galaxies or other objects, many astronomical instruments take images of the wider sky in selected parts of the electromagnetic spectrum using broad or narrow filters. Having the ability to model such images with synthetic data provides an important link in understanding how galaxy properties are connected with their observation.

However, fairly comparing images made from different simulations and models can be problematic, and the size of such datasets poses additional challenges for image generation. Consistency when making images is essential to minimise artificial differences due to the underlying simulation data and processing. With this in mind, TAO was designed to produce mock telescope images in a consistent, seamless, and user friendly way.

The TAO image generation module uses the SkyMaker software package (see Bertin 2009 for further details). This software takes data produced by the light cone and SED modules and creates realistic telescope images that include many of the usual observational effects, like telescope aperture, optical defects, sky characteristics, and aureole around bright galaxies.

TAO users can request multiple images per submitted job. Each image requires the following parameters:

Parameter	Value
Exposure time	8000 sec
Magnitude zero-point	26 mag
Background surface brightness	50 mag arcsec <sup>-2</sup>
Range covered by aureole	50 pixels
Aureole surface brightness	16 mag arcsec <sup>-2</sup>
PSF oversampling factor	7
PSF mask size	512 pixels
Diameter of the primary mirror	3.5 m
Diameter of the secondary mirror	1 m

**Table 2.** SkyMaker default settings in the Mock Image Module initial TAO release. With time these will be added to the module interface for control by the user.

- the desired bandpass filter out of those chosen in the SED module,
- the observed magnitude limits of galaxies to include,
- the redshift range of galaxies to include,
- the opening angles of the image within the light-cone area,
- the RA and Dec coordinates to centre the image on, and
- the resolution of the final output image file.

As input, the Mock Image module requires information about galaxy positions in the light-cone, galaxy morphology, and total magnitudes in the requested filters. From the bulge and the disk properties SkyMaker renders galaxies, using a de Vaucouleurs profile for the bulge and an exponential profile for the disk. Each galaxy is then then placed in the field-of-view to build up the final image.

The Point Spread Function (PSF) model used in SkyMaker is a convolution of the following components: atmospheric blurring for ground-based instruments<sup>5</sup>, telescope motion blurring, instrument diffraction and aberrations, optical diffusion effects, and intra-pixel response. Sky background, noise, saturation, and quantisation modelling are also added by SkyMaker to reflect the inherent noise and artefacts in charge-coupled devices (CCDs).

Some parameters, like image pixel size and average filter wavelength, are calculated by TAO during job processing. Others are pre-chosen to produce the best image quality; in Table 2 we list these. Any further parameters not listed are kept at their default value, as assumed in the SkyMaker software package (Bertin 2009). As the TAO system develops we will add further customisation options to the user interface for more precise image generation control.

## 7 EXAMPLES OF MOCK CATALOGUES

There are many ways in which TAO mock catalogues can be built using the tools described so far. One example is the work of Duffy et al. (2012) to make predictions for the Australian Square Kilometre Array Pathfinder (ASKAP) telescope neutral hydrogen surveys of WALLABY and DINGO. TAO comes with an expanding set of popular survey pre-sets, which can be run as-is or used as a template for more

specific requirements. To illustrate the utility of TAO in this section we outline two common use cases: one to build a representation of the local universe, and the other the universe at higher redshift.

### 7.1 SDSS mock catalogue

One of the largest observational extragalactic databases to date is the Sloan Digital Sky Survey (SDSS) catalogue<sup>6</sup> (Abazajian et al. 2003; Ahn et al. 2012). The SDSS main catalogue covers approximately 14000 square degrees of the sky across a redshift range  $0 < z \lesssim 0.4$ . Many thousands of papers have been written using SDSS data, making it the highest impact repository for extragalactic science in the history of astronomy. Hence, the SDSS is a natural place to start if one wants to construct mock analogues of the local universe.

To reconstruct a SDSS volume using TAO we perform the following steps:

- **General Properties:** We select a light cone geometry, then a simulation and galaxy formation model. Here, the Millennium simulation (Springel et al. 2005) and SAGE galaxy model (Croton et al. in prep.) are adopted. The RA and Dec opening angles are chosen to be 90 and 60 degrees, respectively, with a redshift range of  $0 < z < 0.3$ .
- **Spectral Energy Distribution:** For this particular cone we select a BC03 stellar population model assuming a Chabrier initial mass function. All filters marked with ‘SDSS’ are included for output –  $u, g, r, i, z$  – both apparent and absolute. We apply the Calzetti dust model described in Tonini et al. (2012).
- **Selection:** Our cone is chosen to be volume limited by including all galaxies with stellar masses greater than  $10^8 h^{-1} M_{\odot}$ .

After processing, TAO sends an email to the user with a link to the constructed light cone for download. This particular cone contains 3,122,823 galaxies. In Figure 10 we plot these galaxies to illustrate their spatial distribution. The catalogue also contains as many galaxy properties as predicted by the model and requested by the user. Furthermore, the exact same cone can be reconstructed using a different model in the database, or with a different underlying simulation. This provides the user with an estimate of the theoretical uncertainty between different models and simulations.

### 7.2 CANDELS mock catalogue

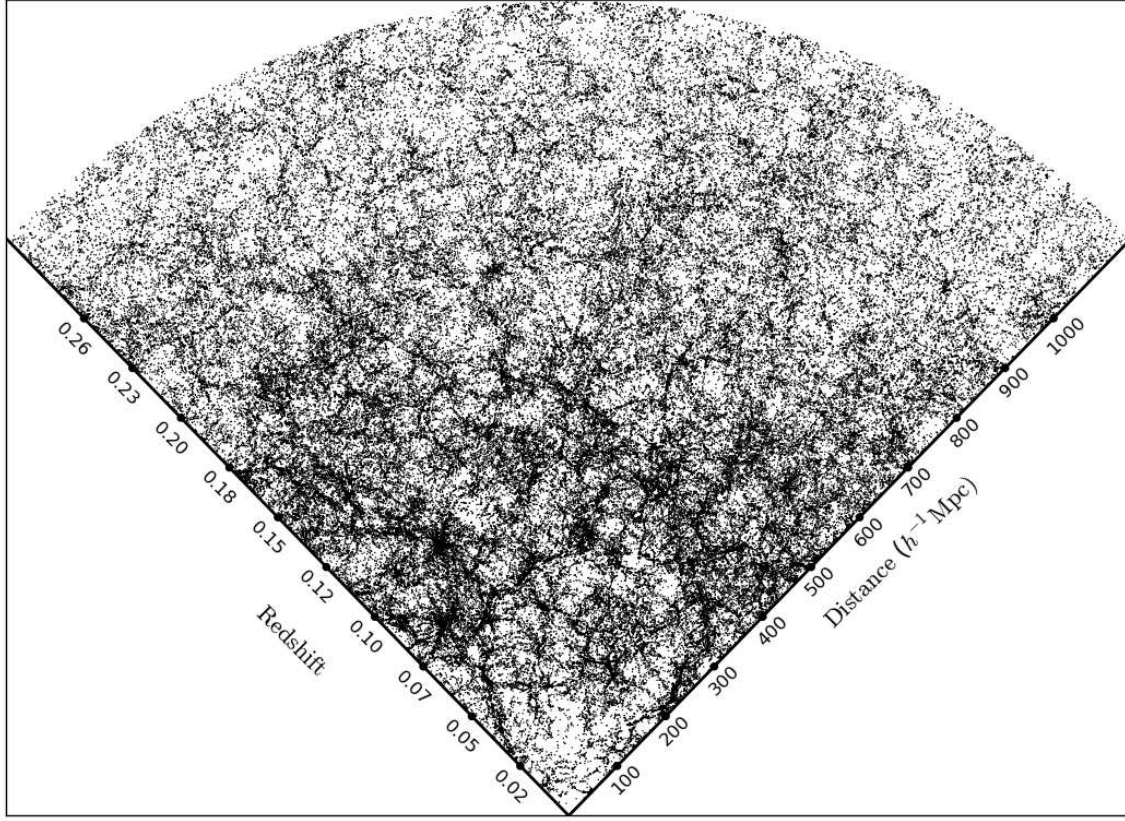
The Cosmic Assembly Near-IR Deep Extragalactic Legacy Survey (CANDELS) is an ongoing Hubble Space Telescope (HST) survey designed to observe galaxy evolution from  $z = 1.5$  to  $\sim 8$  with the WFC3/IR and ACS instruments (Koekemoer et al. 2011; Grogin et al. 2011). It is a useful example with which to explore early galaxy formation using mock catalogues.

Using the TAO facility we built a mock light-cone similar to the CANDELS/WIDE survey:

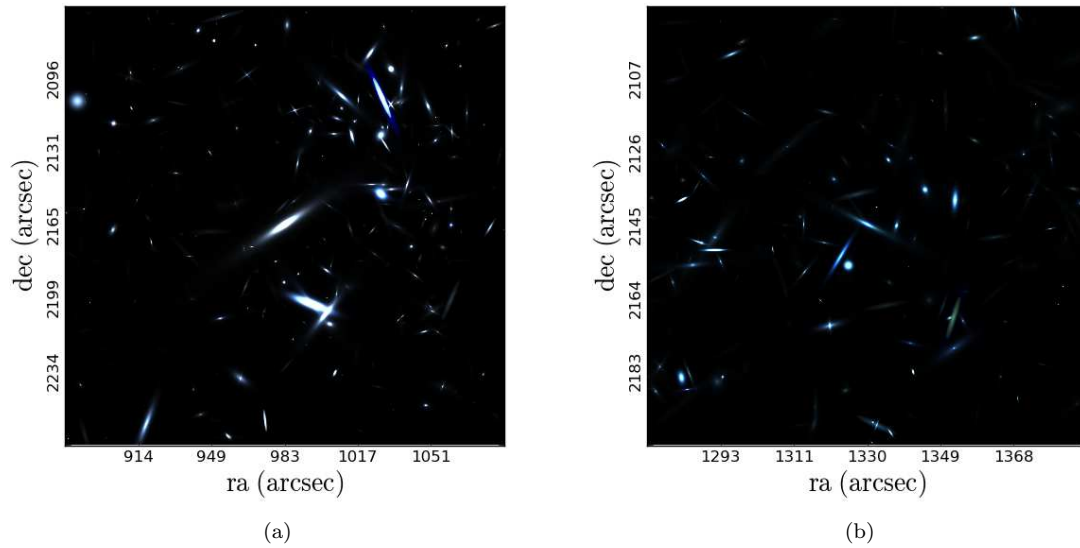
<sup>5</sup> Atmospheric blurring is not currently used in TAO since the default instrument is the Hubble Space Telescope, however ground-based instruments will be added in a subsequent update.

<sup>6</sup> <http://www.sdss3.org/dr9/scope.php>





**Figure 10.** A galaxy mock catalogue with an SDSS limiting magnitude  $r < 17.77$  constructed with TAO.



**Figure 11.** Two composite images of a proto-cluster in TAO using the WFC3 105W, 125W and 160W infra-red filters: a) at  $z = 1$ , with halo mass of  $M_{vir} = 7.3 \times 10^{14} h^{-1} M_{\odot}$  (left), and b) at  $z = 2$ , with a halo mass of  $M_{vir} = 3.6 \times 10^{14} h^{-1} M_{\odot}$  (right).

- **General Properties:** We again select a light cone geometry, this time taking the Bolshoi simulation (Klypin, Trujillo-Gomez & Primack 2011) and SAGE galaxy model (Croton et al. in prep.). Bolshoi is used to take advantage of its higher mass resolution, needed to better resolve lower mass high redshift galaxies. The area on the sky is chosen to be  $1 \times 1$  degrees, with a redshift range of  $0 < z < 9$ .
- **Spectral Energy Distribution:** We select the Conroy, Gunn & White (2009) stellar population model assuming a Kroupa initial mass function. All filters marked with ‘HST/Spitzer’ and ‘HST/WFC3’ are included for output, both apparent and absolute. The slab dust model is used this time.
- **Selection:** Our cone is chosen to be volume limited by including all galaxies with stellar masses greater than  $10^7 h^{-1} M_{\odot}$ .

After downloading the mock catalogue to a local machine we identify the two most massive regions in the cone within a diameter of  $2 h^{-1} \text{Mpc}$  at redshifts  $z = 1$  and  $z = 2$ . We consider these to be proto-cluster environments. The most massive dark matter halo in each field is  $7.3 \times 10^{14} h^{-1} M_{\odot}$  and  $3.6 \times 10^{14} h^{-1} M_{\odot}$ , respectively. The first cluster hosts 90 galaxies, each one more massive than our limit of  $10^8 h^{-1} M_{\odot}$ , while the second contains 40 galaxies. Having identified our objects of interest we re-process the cone with TAO by uploading the previously created XML parameter file and build two mock images centred on each cluster. Both are defined by an imaging area of  $1 \times 1$  arcminutes and cover the redshift ranges  $0.99 < z < 1.01$  and  $1.99 < z < 2.01$ , respectively. For our filter selection we choose the same HST WFC3 filters that CANDELS uses: 105W, 125W, 160W. After processing by TAO the images are then downloaded and combined into a composite which is shown in Figure 11.

## 8 SUMMARY

In this paper we have described a new cloud-based virtual laboratory, the Theoretical Astrophysical Observatory (TAO), that enables any astronomer to produce mock galaxy catalogues based on selectable combinations of a dark matter simulation, semi analytic galaxy formation model, and stellar population synthesis model. TAO is built in a modular fashion, starting with the simulation and model database, that flows upward through a series of science modules to connect to the user via a simple web interface.

The key features of TAO are:

- A flexible design with an intuitive web interface that allows public access to many dark matter simulations, galaxy models, and stellar population synthesis models;
- Advanced techniques to create custom mock light-cones of large volumes with realistic structure;
- Galaxy-by-galaxy spectral energy distribution modelling obtained in post-processing, providing accurate galaxy photometry using galaxy star formation and metallicity histories and positions;
- The ability to image light-cone data using the SkyMaker image processing package, integrated into the TAO workflow.

The flexibility built in to TAO make it useful for many applications. Some examples include:

- Making survey predictions and planning observational strategies;
- The comparison of observational data with simulations and models;
- Testing how different physical prescriptions in the same galaxy model affect galaxy evolution;
- The comparison of different galaxy models run on the *same* dark matter simulation;
- The comparison of a single galaxy model run on *different* dark matter simulations;
- Exploring the effects of different stellar population synthesis models and dust extinction prescriptions on a galaxy’s photometric evolution;

TAO is an open source project and can be freely deployed and further developed by members of the community. The Swinburne TAO portal can also be used as a way to make simulation and galaxy model data available to the public, which is especially useful when the resources to do this in-house are not available or expensive to implement. Enabling the community with free access to state-of-the-art theoretical data facilitates data reuse and multiplies its value.

## ACKNOWLEDGEMENTS

The authors would like to thank Alistair Grant, Jarrod Hurrey, and Gin Tan. We also appreciate the time given by the many astronomers who tested TAO during its initial development. Special thanks goes to Andrew Benson for his help importing the Galacticus semi-analytic model into the TAO database before public release, Volker Springel for making the data from the Millennium simulation publicly available, and Anatoly Klypin for providing access to the Bolshoi simulation. DC acknowledges receipt of a QEII Fellowship by the Australian Research Council (DP1095506). SM, AD and GP are funded from the ARC Laureate Fellowship grant of S. Wyithe (FL110100072).

TAO is part of the All-Sky Virtual Observatory and is funded and supported by Astronomy Australia Limited, Swinburne University of Technology, and the Australian Government. The latter is provided through the Commonwealth’s Education Investment Fund and National Collaborative Research Infrastructure Strategy, particularly the National eResearch Collaboration Tools and Resources (NeCTAR) project. TAO was constructed as a collaboration between Swinburne University of Technology and Intersect Australia and is hosted on the gSTAR national facility at Swinburne.

## REFERENCES

- Abazajian K. et al., 2003, AJ, 126, 2081
- Ahn C. P. et al., 2012, ApJS, 203, 21
- Baugh C. M., 2006, Reports on Progress in Physics, 69, 3101
- Behroozi P. S., Wechsler R. H., Wu H.-Y., 2011, ArXiv e-prints
- Benson A. J., 2012, NA, 17, 175

- Bertin E., 2009, *Mem. Soc. Astron. Italiana*, 80, 422
- Blaizot J., Guiderdoni B., Devriendt J. E. G., Bouchet F. R., Hatton S. J., Stoehr F., 2004, *MNRAS*, 352, 571
- Blaizot J., Wadadekar Y., Guiderdoni B., Colombi S. T., Bertin E., Bouchet F. R., Devriendt J. E. G., Hatton S., 2005, *MNRAS*, 360, 159
- Bower R. G., Benson A. J., Malbon R., Helly J. C., Frenk C. S., Baugh C. M., Cole S., Lacey C. G., 2006, *MNRAS*, 370, 645
- Bruzual G., Charlot S., 2003, *MNRAS*, 344, 1000
- Calzetti D., 1997, *AJ*, 113, 162
- Calzetti D., 2001, *PASP*, 113, 1449
- Carlson J., White M., 2010, *APJS*, 190, 311
- Conroy C., Gunn J. E., White M., 2009, *ApJ*, 699, 486
- Croton D. J., 2013, *PASA*, 30, 52
- Croton D. J. et al., 2006, *MNRAS*, 365, 11
- Daddi E. et al., 2007, *ApJ*, 670, 156
- Davis M., Efstathiou G., Frenk C. S., White S. D. M., 1985, *ApJ*, 292, 371
- De Lucia G., Kauffmann G., White S. D. M., 2004, *MNRAS*, 349, 1101
- Devriendt J. E. G., Guiderdoni B., Sadat R., 1999, *A&A*, 350, 381
- Duffy A. R., Meyer M. J., Staveley-Smith L., Bernyk M., Croton D. J., Koribalski B. S., Gerstmann D., Westerlund S., 2012, *MNRAS*, 426, 3385
- Garel T., Blaizot J., Guiderdoni B., Schaerer D., Verhamme A., Hayes M., 2012, *MNRAS*, 422, 310
- Grogin N. A. et al., 2011, *ApJS*, 197, 35
- Guiderdoni B., Rocca-Volmerange B., 1987, *A&A*, 186, 1
- Hatton S., Devriendt J. E. G., Ninin S., Bouchet F. R., Guiderdoni B., Vibert D., 2003, *MNRAS*, 343, 75
- Henriques B., Maraston C., Monaco P., Fontanot F., Menci N., De Lucia G., Tonini C., 2011, *MNRAS*, 415, 3571
- Hogg D. W., 1999, *ArXiv Astrophysics e-prints*
- Johnston S. et al., 2008, *Experimental Astronomy*, 22, 151
- Keller S. C. et al., 2007, *PASA*, 24, 1
- Kitzbichler M. G., White S. D. M., 2007, *MNRAS*, 376, 2
- Klypin A. A., Trujillo-Gomez S., Primack J., 2011, *ApJ*, 740, 102
- Koekemoer A. M. et al., 2011, *ArXiv e-prints*
- Lemson G., Virgo Consortium t., 2006, *ArXiv Astrophysics e-prints*
- Maraston C., 2005, *MNRAS*, 362, 799
- Mathis J. S., Mezger P. G., Panagia N., 1983, *A&A*, 128, 212
- More S., Kravtsov A. V., Dalal N., Gottlöber S., 2011, *ApJS*, 195, 4
- Overzier R., Lemson G., Angulo R. E., Bertin E., Blaizot J., Henriques B. M. B., Marleau G.-D., White S. D. M., 2012, *ArXiv e-prints*
- Reddy N. A., Steidel C. C., Fadda D., Yan L., Pettini M., Shapley A. E., Erb D. K., Adelberger K. L., 2006, *ApJ*, 644, 792
- Springel V., 2005, *MNRAS*, 364, 1105
- Springel V. et al., 2005, *NAT*, 435, 629
- Springel V., White S. D. M., Tormen G., Kauffmann G., 2001, *MNRAS*, 328, 726
- Tonini C., Bernyk M., Croton D., Maraston C., Thomas D., 2012, *ApJ*, 759, 43
- Tonini C., Maraston C., Devriendt J., Thomas D., Silk J., 2009, *MNRAS*, 396, L36
- Tonini C., Maraston C., Thomas D., Devriendt J., Silk J., 2010, *MNRAS*, 403, 1749
- Tonini C., Maraston C., Ziegler B., Böhm A., Thomas D., Devriendt J., Silk J., 2011, *MNRAS*, 415, 811
- White S. D. M., Frenk C. S., 1991, *APJ*, 379, 52